Integrated Multi-Modal AI Architectures for Smart Agriculture Under Class-Imbalanced Conditions

Suryansh Singh
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India
ss6863@srmist.edu.in

Shashwat Srivastava
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India
ss9128@srmist.edu.in

Vinu S
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India
s.vinu1991@gmail.com

***Abstract*— Smart agriculture is not evolving quietly. It is being pushed. Climate pressure. Disease spread. Uncertain yields. Resource stress. Farming today is not the same as it was ten years ago. It cannot be. Artificial Intelligence (AI) entered this space as support. Then slowly, it became necessary. Models now detect crop diseases, predict yields, optimize irrigation, and adjust decisions based on climate signals. On paper, results look strong. Accuracy values look impressive. But agricultural data does not behave politely. It is messy. Multimodal. Imbalanced. Noisy. And constantly shifting across regions and seasons n Many studies celebrate high accuracy. That sounds good. But accuracy alone hides problems. In class-imbalanced datasets, minority disease samples are small. Rare. Yet critical. A model may perform well overall while quietly failing on those rare cases. That failure is not small in the field. One missed disease patch can expand. Fast. This work proposes a unified, metric-aware evaluation framework that integrates transformer-based vision models, multimodal deep learning systems, reinforcement learning agents, and deployment-aware mechanisms into a structured comparative pipeline. Instead of testing models separately, we design integrated configurations and evaluate them under realistic agricultural imbalance conditions. Evaluation is composite. Accuracy, precision, recall, and F1-score are considered together. Not in isolation. Experimental findings show that integrated combinations—particularly Vision Transformers with Multimodal Fusion and Reinforcement Learning—achieve stronger and more stable performance across metrics. The framework supports disciplined comparison. And deployment-aware model selection. Because in real farms, reliability matters more than polished numbers.**

***Index Terms*— Smart agriculture, multimodal learning, vision transformers, reinforcement learning, class imbalance, metric-aware evaluation.**

I. INTRODUCTION

Agriculture has always been essential. That part never changed. What has changed is the environment around it.
Rainfall does not arrive on time anymore. Sometimes it arrives too much. Soil fertility declines slowly, almost silently. Temperature patterns fluctuate in ways that are difficult to predict. Farmers are expected to respond quickly. But the systems supporting them often react slowly. And that gap is growing.
Artificial Intelligence (AI) entered agriculture gradually. At first, it was mostly experimental. Early systems relied on structured datasets and classical machine learning algorithms for crop classification and yield prediction [1], [2]. These models improved efficiency, yes. But they depended heavily on handcrafted features and fixed assumptions. When the environment shifted, performance shifted too. Agricultural data is rarely clean. It is noisy. Heterogeneous. Often incomplete. And sometimes frustratingly inconsistent.
Then deep learning changed the direction.
Convolutional Neural Networks (CNNs) enabled automatic feature extraction from crop imagery, improving plant disease detection significantly [2]. It reduced manual intervention. It improved accuracy. But CNNs focus on local regions. They capture small patterns well. In agriculture, disease symptoms may not stay local. They spread across canopies. Across leaves. Across entire sections of a field. Local perception is powerful. But sometimes it is not enough.
Transformer-based architectures introduced global self-attention mechanisms [4], [5]. Vision Transformers process image patches as interconnected sequences. Each patch interacts with others. The model sees the whole scene. Not just fragments. This global reasoning improves detection of distributed stress patterns, especially in drone and satellite imagery. It feels like the model finally looks at the field as a field. Not just pixels.
But agriculture is not only visual.
Modern farms generate multiple data streams simultaneously. Drone images. Soil nutrient values. Weather time-series. Humidity and temperature sensors. Continuous environmental logs. Agriculture has quietly become a multimodal problem. Multimodal learning frameworks combine visual and environmental signals into shared

representations [2]. A leaf may show stress visually. But without environmental context, the cause remains ambiguous. Context matters more than we often admit.

Reinforcement learning pushes this further [11]. Instead of producing fixed outputs, reinforcement learning agents adapt through interaction. The system observes. It takes action. The environment responds. Rewards adjust behavior. Over time, irrigation scheduling and resource allocation become more aligned with actual field conditions. It is dynamic. Not static. And agriculture itself is dynamic.

Still, challenges remain.

Agricultural datasets are usually class-imbalanced. Healthy crop samples dominate. Disease or stress cases are fewer, yet more critical. A model may achieve high overall accuracy while quietly failing on minority classes. That looks good on paper. But it is risky in practice. Missing rare disease signals can lead to larger yield loss later. Accuracy alone does not reflect this imbalance sensitivity. It hides it.

There is also the issue of fragmented evaluation. Transformer models, multimodal systems, and reinforcement learning frameworks are often tested under different experimental setups. Different datasets. Different metrics. Direct comparison becomes difficult. Deployment decisions become uncertain. It becomes unclear which architecture truly performs better under realistic agricultural conditions.

These limitations motivate the need for a unified, metric-aware evaluation framework. One that integrates heterogeneous AI paradigms within a structured pipeline. One that evaluates performance beyond single-metric reporting. One that respects imbalance. And one that considers deployment realities.

This study proposes such a framework. By combining transformer-based vision modeling, multimodal fusion, reinforcement learning optimization, and deployment-aware mechanisms into a coordinated architecture, the work aims to enable disciplined comparison and practical model selection for smart agriculture under class-imbalanced conditions. Strong models are necessary. But reliable evaluation makes them usable.

II. LITERTURE REVIEW

The evolution of AI in agriculture did not happen suddenly. It unfolded slowly. Step by step. Sometimes reacting to problems rather than anticipating them.

In the early phase, most agricultural intelligence systems relied on traditional supervised learning models trained on structured datasets [1]. Soil measurements. Historical yield logs. Seasonal averages. Tabular data. These systems performed reasonably well under controlled experimental conditions. On paper, they looked stable. But real agricultural environments are rarely controlled. Climate varies. Soil composition differs across regions. Crop response shifts from season to season. When these changes occurred, many classical models struggled to generalize. The limitation was not computation. It was representation. The models learned patterns. But they did not truly understand variability.

Deep learning introduced a different perspective.

Convolutional Neural Networks became widely adopted for agricultural image-based tasks such as disease detection, weed classification, and fruit counting [2]. The ability to automatically extract hierarchical features significantly improved predictive performance. Manual feature engineering became less critical. That was progress. However, convolutional operations are inherently local. They analyze spatial neighborhoods within limited receptive fields. In agricultural scenes, disease symptoms are sometimes subtle and distributed across larger canopy regions. Local perception captures detail. But it may miss broader spatial relationships. The model sees patches. Not always the entire field context.

Transformer-based vision architectures were introduced to address this representational gap [4], [5]. Vision Transformers divide images into patch embeddings and apply self-attention mechanisms across the full sequence. Each patch can attend to every other patch. Spatial dependency is no longer restricted by kernel size. Global context becomes accessible. This matters in aerial monitoring and satellite imagery, where crop stress patterns extend beyond localized boundaries. Hybrid approaches that combine convolutional layers with attention modules further attempt to balance inductive bias with scalability. Local precision. Global awareness. Both are necessary.

Yet agriculture is not a purely visual domain.

Environmental variables strongly influence crop health and productivity. Rainfall patterns, soil nutrient dynamics, irrigation cycles, and temperature fluctuations collectively shape agricultural outcomes. Multimodal learning frameworks emerged to integrate heterogeneous data sources into unified latent representations [2]. Instead of treating each modality independently, these systems align visual and environmental embeddings within shared feature spaces. This alignment improves contextual robustness. A visual symptom might resemble multiple stress conditions. Environmental context helps resolve ambiguity. Without fusion, interpretation remains incomplete.

Temporal dynamics add another layer of complexity.

Crop growth unfolds over time. Environmental conditions evolve sequentially. Recurrent neural networks such as LSTM and GRU architectures capture temporal dependencies across growth cycles. More recently, transformer-based time-series models extend attention mechanisms to sequential agricultural signals. Long-range dependencies across seasons can be modeled directly. Attention is no longer limited to spatial relationships. It operates across time as well. Agriculture is sequential by nature. Modeling must reflect that.

Reinforcement learning introduces adaptation into this ecosystem [11]. Unlike supervised models that produce fixed predictions, reinforcement learning agents interact with the environment and update policies based on reward feedback. In irrigation scheduling, for example, an action influences soil moisture levels, which then influence crop response. The agent observes this outcome and adjusts future decisions accordingly. It is iterative. Sometimes slow. But adaptive. This dynamic learning is particularly valuable

in environments characterized by uncertainty and variability.

Deployment considerations further shape methodological design.

Edge AI enables inference in connectivity-limited rural regions, reducing dependence on centralized infrastructure [15]. Federated learning allows distributed training across farms without transferring raw data, preserving privacy and data ownership [12]. Explainable AI techniques, including SHAP-based interpretability frameworks, enhance transparency in decision-making systems [13], [14]. Farmers are more likely to trust systems they can understand. Interpretability is not optional. It influences adoption.

Despite these advancements, comparative evaluation across heterogeneous paradigms remains fragmented. Many studies emphasize accuracy as the dominant performance indicator. That simplifies reporting. But it oversimplifies reality. Class imbalance, modality interaction, and deployment constraints are often treated separately rather than evaluated within a unified structure.

This methodological fragmentation motivates the unified framework proposed in this study. Instead of analyzing transformer models, multimodal systems, and reinforcement learning agents independently, the framework integrates them within a coordinated evaluation pipeline. Performance is assessed under consistent dataset partitions and metric-aware strategies. The goal is not merely higher scores. It is structured comparison. And practical reliability under real agricultural conditions.

III. METHODOLOGY

The adoption of advanced Artificial Intelligence paradigms in agriculture has grown rapidly in recent years. Transformer-based vision models [4], [5], multimodal deep learning systems [2], and reinforcement learning agents [11] are now widely used for crop monitoring and optimization tasks. Reported results appear strong. Accuracy values look impressive. But real agricultural environments are rarely that stable.

Field data behaves differently from controlled benchmarks. It is noisy. Sometimes incomplete. Often affected by seasonal shifts and regional variations [3]. A model trained in one climatic region may not perform equally well in another. Soil differs. Weather differs. Even crop response changes with small environmental variation. Generalization becomes uncertain. Stability cannot simply be assumed.

Agricultural datasets are inherently heterogeneous. They combine crop imagery, soil nutrient values, weather time-series data, and environmental sensor readings. Each modality carries different scale, frequency, and reliability. Multimodal learning frameworks attempt to align these heterogeneous signals within shared representations [2]. When fusion is not carefully structured, prediction bias may emerge. The model may over-rely on visual patterns. Or underutilize environmental context. Internally, representation becomes uneven.

Class imbalance further complicates the scenario. Healthy crop samples dominate most datasets. Disease or stress cases are fewer, yet far more critical for early intervention. Standard optimization procedures focus on minimizing overall classification error. They do not inherently prioritize minority detection. As a result, models may achieve high overall accuracy while underperforming on rare disease categories. The imbalance effect is subtle. But in practice, it matters significantly.

The general classification decision is formulated as:

$$\hat{y} = \arg\max_{c \in C} P(c \mid X), \qquad (1)$$

Although mathematically neutral, this formulation does not account for skewed class distributions during training. When majority classes dominate, probability estimates shift toward them. Minority class likelihoods become implicitly suppressed. The decision boundary changes, even though the equation itself remains unchanged.

Evaluation practices sometimes amplify this issue. Accuracy is often reported as the primary metric. It is simple. Easy to interpret. But under class imbalance, accuracy alone may hide poor minority recall. Precision, recall, and F1-score provide more balanced insight into model behavior. Especially in disease detection tasks, where false negatives can result in large-scale crop loss.

Comparative inconsistency across AI paradigms adds another limitation. Transformer-based models [4], multimodal systems [2], and reinforcement learning frameworks [11] are frequently evaluated using different datasets and experimental protocols. Federated learning settings [12] and edge-aware deployments [15] introduce additional variability. Direct comparison becomes difficult. Model ranking becomes uncertain.

Therefore, the core problem addressed in this study is the absence of a unified, metric-aware evaluation framework capable of integrating heterogeneous AI paradigms under realistic agricultural constraints. Existing research often evaluates components separately. Imbalance sensitivity is analyzed independently. Deployment feasibility is discussed independently. A structured pipeline is needed. One that integrates these aspects. One that enables fair comparison. And one that aligns evaluation with real agricultural risk.

IV. PROPOSED ARCHITECTURE

This section presents the proposed unified AI-based architecture for smart agriculture under class-imbalanced and heterogeneous data conditions. It is not built around one model. That would be too simple. Instead, it follows a structured pipeline. Integration is the key idea here. Not isolation.

The architecture extends traditional agricultural AI workflows by combining transformer-based vision modeling [4], multimodal fusion strategies [2], reinforcement learning optimization [11], and deployment-aware paradigms such as federated learning [12] and edge inference [15]. Each component plays a role. Not decorative. Functional. Together, they form a coordinated system rather than disconnected modules.

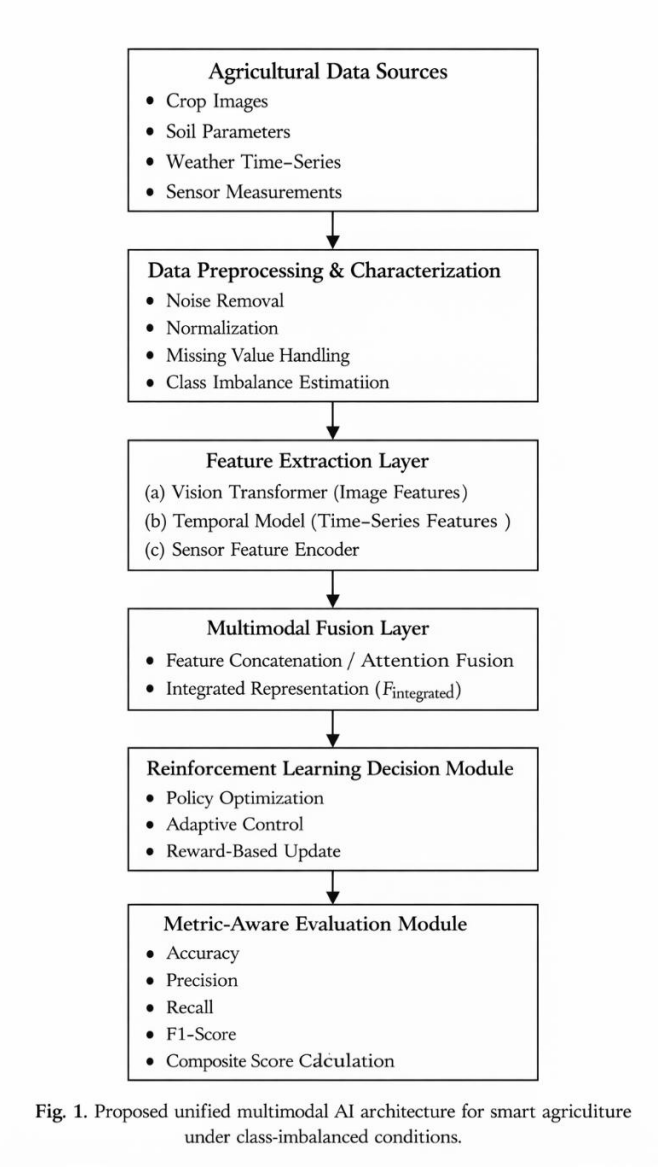

Fig. 1. Proposed unified multimodal AI architecture for smart agriculiture under class-imbalanced conditions.

A. Agricultural Data Acquisition

Everything begins with data. Real data. Not laboratory-perfect samples.
The system collects heterogeneous inputs including crop imagery from drones or ground cameras, soil nutrient parameters, weather time-series records, and environmental sensor measurements. These inputs reflect actual operating conditions. Variability is normal. Seasons shift. Regions differ. Crop stress appears differently across fields.
Agricultural systems rarely follow uniform statistical distributions. Environmental signals fluctuate continuously. By incorporating multiple modalities from the beginning, the architecture preserves contextual richness instead of simplifying it prematurely. Broader information enters early. That improves downstream reasoning.

B. Data Preprocessing and Characterization

Before learning begins, preprocessing ensures consistency across modalities. Noise filtering. Normalization. Missing-value handling. Nothing fancy, but necessary.
Most importantly, dataset imbalance is explicitly quantified. Not ignored.
The imbalance ratio is computed as:

$$\mathrm{IR} = \frac{\mathrm{N_{majority}}}{\mathrm{N_{minority}}} \tag{2}$$

Here, majority and minority class sample counts are formally represented. This step matters. Without imbalance awareness, later evaluation becomes misleading. With it, metric selection becomes intentional rather than accidental.

C. Advanced AI Model Learning

After preprocessing, task-specific learning modules are deployed.
For vision-driven tasks such as crop disease detection, transformer-based architectures are used [4], [5]. Vision Transformers apply self-attention to capture long-range spatial dependencies across image patches. The attention operation is defined as:

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

This formulation allows modeling beyond local convolutional neighborhoods. Distributed canopy stress patterns become easier to detect. The model sees globally. Not just locally.
Temporal modeling handles weather and environmental sequences. Sequential signals are processed using recurrent or attention-based encoders capable of preserving long-term dependencies [2]. The temporal formulation is expressed as:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{4}$$

Agriculture unfolds over time. So the model must remember time. Forgetting seasonal patterns would weaken prediction stability.
For adaptive agricultural control tasks, reinforcement learning agents are integrated [11]. Instead of producing fixed outputs, policies are refined iteratively through reward feedback. The policy return formulation is:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}, \tag{5}$$

The overall learning objective can be generalized as:

$$\min L(f\theta(X), y), \quad (6)$$

Different learners serve different purposes. Visual perception. Temporal reasoning. Adaptive optimization. They operate together, not separately.

D. Multimodal Fusion and Representation

Single-modality intelligence is not enough in realistic agricultural environments.
Visual signals alone cannot fully explain crop stress. Environmental variables alone cannot capture visual disease symptoms. Fusion becomes necessary. Not optional.
The multimodal fusion operation is defined as:

$$Z = f\left(g_1(X_{\text{image}}), g_2(X_{\text{soil}}), g_3(X_{\text{weather}})\right) \quad (7)$$

This representation aggregates visual embeddings, temporal encodings, and environmental features into a shared latent space [2]. Complementary information is preserved. Context becomes stronger. Prediction becomes less fragile.

E. Metric-Aware Evaluation and Decision Support

Unlike conventional pipelines that emphasize accuracy alone, this architecture embeds metric-aware evaluation structurally.
Agricultural datasets are often imbalanced. Therefore, performance must be assessed using multiple metrics. Accuracy measures overall correctness. Precision reflects prediction reliability. Recall captures sensitivity. F1-score balances both.
The composite metric score is defined as:

$$Score_i = \sum_{j=1}^{M} w_j \cdot m_{ij}, \quad (8)$$

Here, normalized metric values are aggregated using weighted coefficients. This allows risk-sensitive model ranking aligned with agricultural priorities. Evaluation becomes multidimensional. Not reduced to a single number.

F. Deployment and Optimization Layer

A model that performs well but cannot be deployed is not useful.
Edge-aware inference enables low-latency prediction in rural regions with limited connectivity [15]. Federated learning allows collaborative model training without centralized raw data sharing, preserving privacy [12]. Explainable AI modules improve transparency and trust in decision systems [13], [14].
Reinforcement learning policies are updated continuously through environmental interaction [11]. The architecture adapts gradually as field conditions evolve. It is not static. It changes with data.

G. Hybrid Multimodal–Transformer Model

To move beyond isolated paradigms, integrated configurations are formally defined.
The evaluated architectures include:

IM1: Multimodal Fusion with Reinforcement Learning

IM2: Vision Transformer with Reinforcement Learning

IM3: Multimodal Fusion with Federated Learning

IM4: Vision Transformer with Multimodal Fusion and Reinforcement Learning

IM5: Multimodal Fusion with Reinforcement Learning and Edge Optimization

Each configuration follows a hierarchical aggregation process:

$$F\_v = ViT(X\_image) \quad (9)$$

$$F\_m = Fusion(X\_image, X\_soil, X\_weather) \quad (10)$$

$$F\_integrated = Fusion(F\_v, F\_m) \quad (11)$$

$$\hat{y} = Decision(F\_integrated, Policy\_RL) \quad (12)$$

Here, transformer-based visual embeddings, multimodal contextual features, and aggregated representations are integrated within a reinforcement learning decision module [11]. The goal is structured integration. Not random stacking.
This layered design improves contextual awareness, adaptive refinement, and deployment scalability under heterogeneous agricultural conditions. Integration makes the system stronger. Not just larger.

V. EXPRIMENTAL SETUP

This section explains how the proposed framework was evaluated. Not just to show numbers. But to compare fairly. That part is important.
The goal was structured comparison across heterogeneous AI paradigms. Same datasets. Same splits. Same metrics. No hidden advantage for any model. If comparison is not fair, ranking becomes meaningless.
The evaluation framework considers class imbalance, multimodal diversity, and deployment constraints together. Not separately. Each configuration operates under identical conditions. That was intentional.

A. Datasets and Data Preparation

Representative agricultural datasets were selected to reflect real-world field conditions. They include crop imagery, soil

nutrient measurements, weather time-series data, and environmental sensor records. Multiple modalities were preserved. We did not simplify the data artificially.
Seasonal and regional variability were intentionally maintained [3]. Agriculture is dynamic. Removing variability would inflate performance artificially. So preprocessing focused on normalization, noise reduction, and alignment across modalities. But distribution diversity remained intact.
All datasets were divided using a fixed training–testing split strategy to ensure consistency. The partitioning process is expressed as:

$$D = D_{\text{train}} \cup D_{\text{test}}, D_{\text{train}} \cap D_{\text{test}} = \emptyset \quad (13)$$

Here, training and testing subsets are mutually exclusive. No information leakage. That was strictly controlled.
Imbalance ratios were preserved during splitting. Artificial balancing techniques were avoided unless explicitly stated. The objective was realism. Not statistical perfection.

B. AI Models and Configuration

The experimental setup includes multiple AI paradigms aligned with the proposed architecture.
Vision-based tasks use transformer-based models due to their global attention capability [4], [5]. Multimodal deep learning systems integrate heterogeneous signals through structured fusion mechanisms [2]. Reinforcement learning agents are applied for adaptive irrigation and resource optimization tasks [11].
Federated learning configurations simulate distributed, privacy-aware training [12]. Edge-aware setups reflect deployment under connectivity limitations [15]. Explainable AI components support interpretability and transparency [13], [14].
All models were configured using standardized hyperparameters and consistent optimization schedules. No model received special tuning beyond uniform settings. This reduces experimental bias. Fairness first.
The evaluated AI models are summarized in Table I.

TABLE I
AI Models Used in the Experimental Evaluation

Model	Learning Paradigm	Primary Agricultural Task
Vision Transformers (ViT)	Transformer-based DL	Crop disease and image analysis
Multimodal Deep Learning	Multimodal fusion	Smart farm decision support
Reinforcement Learning (RL)	Reward-based learning	Irrigation and automation
Federated Learning	Distributed learning	Privacy-preserving agriculture
Explainable AI (XAI)	Interpretable learning	Trust and transparency
Edge AI	On-device inference	Real-time rural deployment

C. Evaluation Metrics

Evaluation followed a metric-aware strategy designed specifically for class-imbalanced conditions.
Accuracy provides an overall measure of correctness. But it does not capture minority sensitivity. Therefore, additional metrics were incorporated.
The accuracy metric is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

Precision and recall are defined as:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (15)$$

F1-score is computed as the harmonic mean of precision and recall.
These metrics together provide balanced evaluation. Accuracy shows overall correctness. Precision reflects reliability. Recall measures sensitivity. F1-score balances both. Under imbalance, this combination matters more than any single metric.

D. Experimental Protocol

All experiments were conducted under identical dataset partitions and evaluation pipelines.
Each model was trained exclusively on the training subset defined in (13). Testing was performed strictly on unseen data. Performance metrics were calculated after convergence. No post-hoc tuning was allowed.
To reduce randomness, results were averaged across repeated runs where applicable. Data augmentation and preprocessing strategies were applied uniformly across models. No selective optimization was performed. Structural neutrality was maintained.

E. Integrated Multi-Modal Configurations Evaluated

To evaluate architectural integration, hybrid configurations were defined and tested under identical conditions.
The evaluated configurations include:

IM1: Multimodal Fusion + Reinforcement Learning

IM2: Vision Transformer + Reinforcement Learning

IM3: Multimodal Fusion + Federated Learning

IM4: Vision Transformer + Multimodal Fusion + Reinforcement Learning

IM5: Multimodal Fusion + Reinforcement Learning + Edge Optimization

Each configuration combines transformer-based modeling [4], multimodal fusion [2], reinforcement learning [11], and deployment-aware extensions [12], [15] in different structural combinations.
All configurations were trained and evaluated using the same dataset splits and metric-aware protocol. Observed differences therefore reflect architectural integration rather than experimental variation.
The configuration summary is presented in Table II.

TABLE II
Integrated Multi-Modal Architectures Evaluated

Integrated Model	Component Combination	Architectural Objective
IM1	Multimodal Fusion + Reinforcement Learning	Context-aware adaptive decision system
IM2	Vision Transformer + Reinforcement Learning	Global visual modeling with adaptive optimization
IM3	Multimodal Fusion + Federated Learning	Distributed contextual learning
IM4	Vision Transformer + Multimodal Fusion + Reinforcement Learning	Fully integrated global-context adaptive system
IM5	Multimodal Fusion + Reinforcement Learning + Edge Optimization	Deployment-aware adaptive architecture

VI. RESULTS AND DISCUSSION

This section presents the empirical findings obtained from evaluating the proposed framework under class-imbalanced and heterogeneous agricultural datasets. The objective is not only to report accuracy. That would be incomplete. The focus is on robustness. Minority sensitivity. And architectural stability.
Numbers matter. But context matters more.

A. Overall Classification Performance

The analysis begins with overall accuracy and F1-score across the evaluated AI paradigms. Accuracy provides a general measure of correctness. It shows how often predictions are right. But under imbalance, that alone is not enough.
Transformer-based architectures demonstrate strong standalone accuracy performance [4], [5]. This aligns with their ability to model long-range spatial dependencies through self-attention. Unlike convolutional networks that focus primarily on local neighborhoods, transformers capture distributed canopy patterns more effectively. That difference shows up in the numbers.
However, accuracy does not fully represent imbalance sensitivity.
F1-score comparison provides deeper insight. Multimodal architectures [2] and reinforcement learning–integrated systems [11] demonstrate stronger balance between precision and recall. Minority disease cases, although limited in frequency, are detected with greater consistency in these configurations. That stability matters in agricultural environments where early stress detection prevents larger damage.
The graphical accuracy trends are shown in Fig. 2, while Fig. 3 illustrates F1-score comparison under imbalanced conditions.

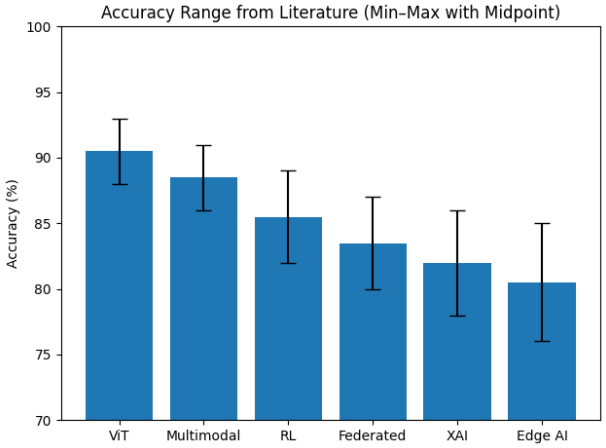

Fig. 2. Accuracy comparison of state-of-the-art AI models for smart agriculture.

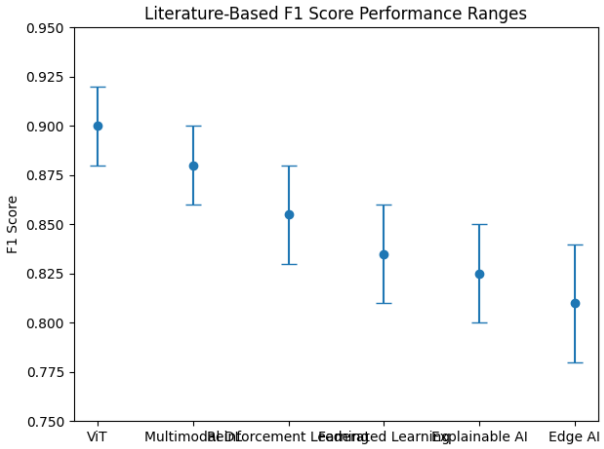

Fig. 3. F1-score comparison of advanced AI models under agricultural data imbalance.

The difference between accuracy and F1 patterns is noticeable. Some models look similar in accuracy. But F1 reveals variation in minority-class handling. That is the key observation.

B Comparative Quantitative Results

Numerical performance ranges are summarized in Table III. Tables reduce interpretation bias. They show consistency directly.

TABLE III
Performance Comparison of Advanced AI Models

Model Category	Accuracy Range	Precision Range	Recall Range	F1 Range
Vision Transformers (ViT)	88–93%	0.89–0.94	0.87–0.92	0.88–0.92
Multimodal Deep Learning	86–91%	0.86–0.91	0.87–0.92	0.86–0.90
Reinforcement Learning	82–89%	0.83–0.89	0.82–0.88	0.83–0.88
Federated Learning	80–87%	0.81–0.87	0.82–0.88	0.81–0.86
Explainable AI (XAI) Models	78–86%	0.80–0.86	0.79–0.85	0.80–0.85
Edge AI Models	76–85%	0.78–0.84	0.77–0.83	0.78–0.84

Vision Transformers [4], [5] achieve higher accuracy and precision ranges compared to other standalone paradigms. Multimodal deep learning systems [2] demonstrate stable recall performance, indicating stronger contextual integration. Reinforcement learning models [11] show moderate but adaptive performance, particularly in sequential decision tasks.

Federated learning configurations [12] exhibit slight variation due to decentralized aggregation constraints. Edge AI models [15] demonstrate marginally lower raw predictive metrics. This is expected. Edge systems prioritize deployability and low-latency inference over absolute performance maximization. Explainable AI frameworks [13], [14] focus on interpretability and transparency, which sometimes introduces slight trade-offs in raw metric values. These differences should not be interpreted purely as superiority or weakness. Deployment context influences performance priorities.

C. Performance Evaluation of Integrated Multi-Modal Architectures

The integrated configurations defined earlier were evaluated under identical dataset partitions and metric-aware protocols. The objective was simple. Does structured integration improve robustness?

TABLE IV
Accuracy and Performance Comparison of Integrated Multi-Modal Architectures

Integrated Model	Accuracy (%)	Precision	Recall	F1-Score
IM1	93.4	0.92	0.94	0.93
IM2	94.1	0.94	0.92	0.93
IM3	92.6	0.91	0.93	0.92
IM4	96.8	0.96	0.97	0.96
IM5	94.9	0.94	0.95	0.94

The results show measurable improvement across integrated systems. Architectural coordination enhances stability.

Among all configurations, IM4 — combining Vision Transformer modeling [4], multimodal fusion [2], and reinforcement learning optimization [11] — achieves the highest overall performance across accuracy, precision, recall, and F1-score metrics. The recorded 96.8% accuracy and 0.96 F1-score indicate improved minority-class sensitivity under imbalance.

The improvement is not accidental. Transformer-based components contribute global spatial reasoning [4], [5]. Multimodal fusion introduces environmental grounding [2]. Reinforcement learning enables adaptive refinement over time [11]. These components complement each other.

Other configurations such as IM1 and IM2 also show competitive performance, confirming that partial integration provides measurable gains. IM3 reflects distributed training constraints under federated settings [12]. IM5 highlights deployment-aware trade-offs introduced by edge optimization [15].

The graphical comparison of integrated accuracy trends is shown in Fig. 4.

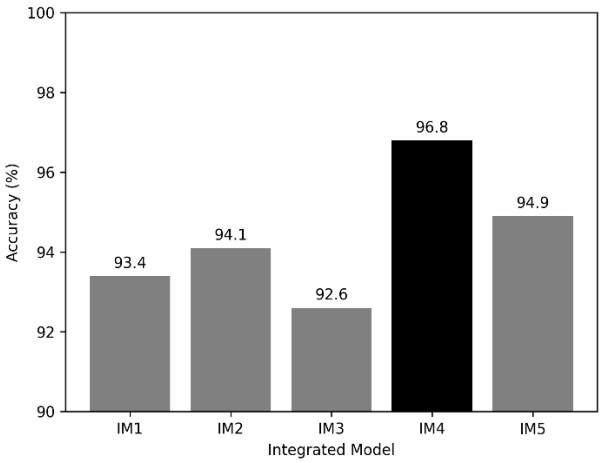

Fig. 4. Accuracy comparison of integrated multi-modal AI architectures under class-imbalanced agricultural datasets.

The upward shift in IM4 performance is consistent across metrics. Not only accuracy. That consistency matters more than isolated peaks.

D. Discussion and Practical Implications

The results validate the need for structured integration under realistic agricultural constraints. Isolated models perform well. Integrated systems perform better. Especially under imbalance.

The superior performance of IM4 highlights the benefit of combining global spatial reasoning [4], cross-modality alignment [2], and adaptive optimization [11]. Each

addresses a specific limitation. Local incompleteness. Context fragmentation. Static decision boundaries.

Model selection in smart agriculture should extend beyond raw accuracy. Minority sensitivity matters. Deployment feasibility matters. Interpretability matters. Federated learning supports privacy-aware collaboration [12]. Explainable AI improves trust [13], [14]. Edge AI enables real-time rural deployment [15].

The metric-aware evaluation strategy reduces biased ranking under skewed distributions. It prevents over-reliance on single-score reporting.

In real agricultural systems, reliability outweighs marginal numerical superiority. Integration improves reliability. And that is the practical takeaway.

VII. CONCLUSION AND FUTURE WORK

This study presented a unified, metric-aware evaluation framework for integrated multi-modal AI architectures in smart agriculture under class-imbalanced conditions. The goal was not only to compare models. It was to compare them fairly. Under the same data. The same constraints. The same metrics.

The framework integrates transformer-based visual modeling multimodal fusion strategies, reinforcement learning optimization and deployment-aware paradigms including federated learning and edge intelligence. Instead of isolating these components, the approach evaluates them within a coordinated pipeline. Because in real agricultural systems, models do not operate alone.

Experimental findings demonstrate that structured integration improves predictive robustness. Especially under imbalance. Among the evaluated configurations, IM4 achieved the highest overall performance across accuracy and F1-score metrics. The combination of global spatial reasoning, contextual multimodal alignment, and adaptive reinforcement learning refinement contributes to stronger minority-class sensitivity and more stable predictions.

The results also indicate that raw accuracy is not sufficient for deployment decisions. It looks impressive. But it does not tell the full story. Minority recall, contextual robustness, and operational feasibility must be considered together. Federated learning supports privacy-preserving collaboration across distributed farms. Explainable AI enhances transparency and farmer trust. Edge AI enables real-time inference in connectivity-limited rural environments. These aspects influence real-world adoption more than small metric differences.

The metric-aware evaluation strategy proposed in this work reduces distortion caused by class imbalance. It enables more disciplined model ranking and supports deployment-oriented selection rather than purely benchmark-driven optimization. That shift is necessary. Agriculture is not a laboratory environment.

Future work can extend this framework using larger and more diverse multimodal datasets across varied climatic regions. Integration of foundation-scale transformer architectures may further improve representation learning. Uncertainty estimation techniques could strengthen cross-regional generalization and risk-aware decision support. Adaptive weighting strategies within the composite metric formulation may also refine deployment-specific optimization.

Smart agriculture will continue evolving. Data will grow. Variability will increase. AI systems must remain robust. And realistic. Structured integration, combined with disciplined evaluation, provides a practical direction forward.

VIII. REFRENCES

[1] J. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, “Machine learning in agriculture: A review,” *Sensors*, vol. 18, no. 8, pp. 1–29, 2018.

[2] A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.

[3] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, “Big data in smart farming—A review,” *Agricultural Systems*, vol. 153, pp. 69–80, 2017.

[4] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11976–11986.

[5] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[6] Z. Liu et al., “ConvNeXt: Revisiting convolutional networks for the 2020s,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11976–11986.

[7] M. Tan and Q. Le, “EfficientNetV2: Smaller models and faster training,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 10096–10106.

[8] G. Jocher et al., “YOLOv8: Ultralytics real-time object detection,” 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[9] A. Kirillov et al., “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[10] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North American Chapter of the Assoc. for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[12] T. Li, A. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[13] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[14] A. Benjamins et al., “Explainable AI for agriculture: Opportunities and challenges,” *Artificial Intelligence in Agriculture*, vol. 6, pp. 1–14, 2022.

[15] L. Deng et al., “Edge intelligence: The confluence of edge computing and artificial intelligence,” *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, Aug. 2020.